

周报总结

2015-07-12

前周回顾：

为了试验之前我们的方法，能否对部分聚类进行展示，所以我修改了聚类方法，将之前的统一聚类改成了部分聚类，但是由于之前系统并没有针对部分聚类结果进行展示，因此，要兼容部分展示，还要对系统进行一定程度上的修改。比如现在只能看转移，还不能看具体其中的类别数据分布，这是因为，我们之前进行展示分布的时候，为了提高展示的效率，我们数据采用了预处理+实时计算的方法，部分数据预处理，部分数据实时计算，这也导致，要想兼容部分属性的展示，就必须修改系统（系统程序过去对数据的耦合很强）。

本周总结：

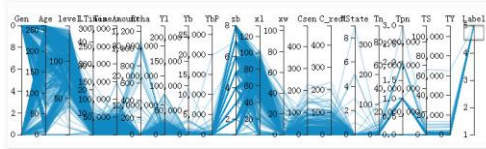
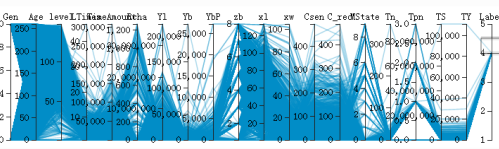
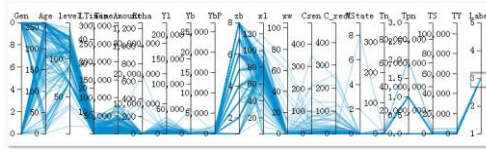
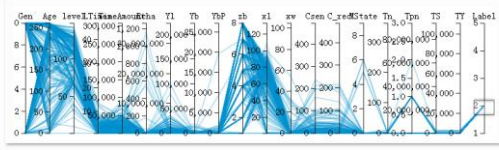
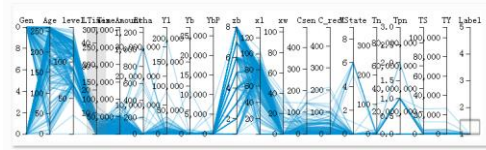
本周开始的是后在对系统进行修改，但是由于我们所有的绘制，可视化方法并没使用别人的工具和库，全部是自己写的，而我们想要尽快测试一下方法，因此，巫老师建议，希望能用现成的 JS 工具 D3 对数据进行简单的测试性质的展示，主要用的是平行坐标的方式。

自己因为之前对 D3 和 WEB 端不熟，所以稍花了一点时间，在配置 Apache&PHP&Mysql 上，使用的是 Xampp 工具套件，在开始安装的时候，各种报错不知道原因在哪里，后来查了半天资料，发现是我现在使用虚拟机还有一些应用程序的端口占用了服务器的端口。最后，请教了下万祺一起研究了，把问题解决了。

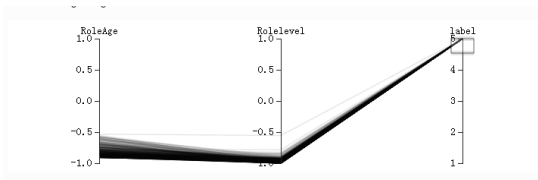
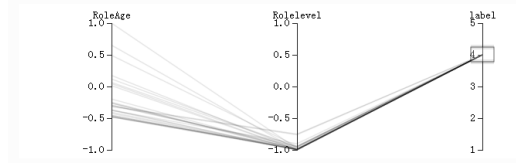
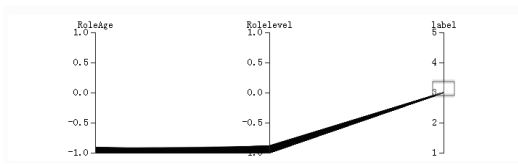
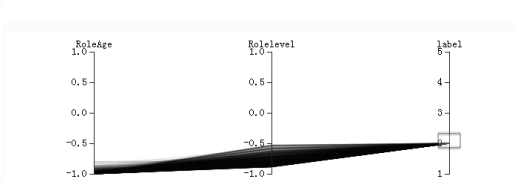
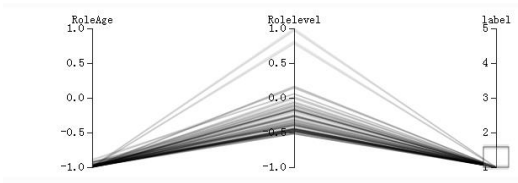
然后是使用 D3 的问题，自己之前几乎没有写过 JS，只是在本科做软件安全方向调试浏览器漏洞的时候，简单的了解了一下。D3 在使用的时候存在了很多问题，比如说：view 不能加载，比如说，加载出来的组件颜色没有区分，标题顺序有问题等等。自己半是研究，半是询问小组里用 JS 的同学。绝大部分有所解决。

具体工作

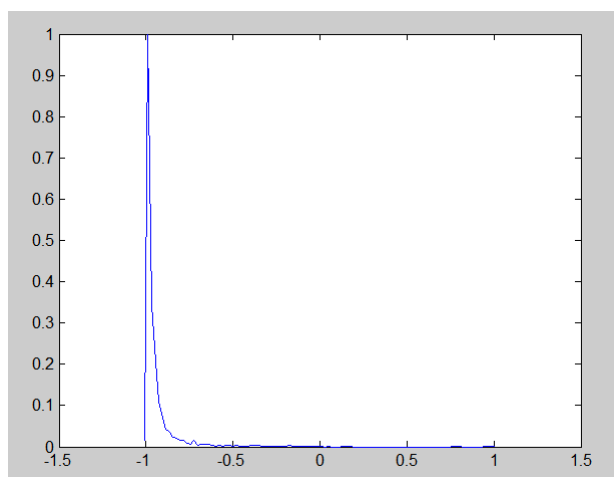
工作主题主要是对部分聚类好的数据，利用平行坐标进行展示。下图为最开始做出的效果：



使用的部分聚类数据为 5-9 标签的数据，数据比较难看出变化很多都聚在了一起，比较费解，以为是聚类算法的问题。然后尝试测试，使用比较少的类别聚类测试一下：



可以看出 采用较小的数据聚类的时候还是能看出一些不同的。但是图 2 图 3 和图 5 其实区分的并不是特别的明显。而图 1 图 4 虽然聚出了类但是 数据量比较小。对此感觉十分的奇怪，然后仔细想了一下。突然想到是不是数据的问题。于是计算了一下数据的概率分布情况，如下：



数据是经过归一化处理过的，从-1 到 1 之间。计算发现，90%的数据集中在整个数据空间的6%的空间，99%的数据聚集在整个数据空间的的前30%的空间，而99.9%的数据聚集在40%左右的空间，可见我们的数据有极少一部分，值很大的点。这些点的数目很少，然而因为这些点的存在，导致了我们的数据归一化的时候绝大部分的数据其实是距离很近，然而又不是所有的数据都存在这种情况，有一些数据分布还是比较均匀的，因此直接对整个聚类造成了影响。影响不但体现在聚类上，同时也体现在最终结果展示上，因为有时我们看到的那些线，似乎没有变化，大家是一样的，但其实是有区别的只是他们堆在了一起 距离很近。

因为之前做聚类的时候，因为比较简单有现成的工具，因此，处理数据和做聚类部分其实是交给数院的学弟学妹们做的，自己没有亲自去看数据分布和他们做的效果。所以并没有发现这个数据有孤立点的情况。

下周安排：

- 1 处理数据：并非所有的数据都能用到，在对数据进行处理的时候，要有一定保留和筛选，计划使用的筛选策略是计算数据的概率分布，将那些过大的点（阈值选取99%的数据集中的范围）筛掉。
- 2 如果效果仍然不好，尝试换一个聚类算法。